Stochastické simulačné metódy (2-PMS-123)
Úlohy na samostatné vypracovanie

Radoslav Harman, KAMŠ, FMFI UK

10. októbra 2020

Každý z Vás si zvolí jedno zadanie. Riešenie očakávam vo forme programu pre `R`, `Matlab` alebo `Python`, ktorý mi pošlete na môj email pred začatím skúšky. V prípade, že použijete `Matlab` alebo `Python`, doneste si prosím vlastný notebook s funkčnou inštaláciou použitého prostredia.

Píšte prehľadný, komentovaný kód, s konzistentnou logikou značenia, napríklad v štýle tidyverse. Program vypracujte tak, aby Vaše simulačné výsledky mohol ktokoľvek presne replikovať. Vypracovaný program si dôkladne prekontrolujte na viacerých, pokiaľ možno čo najpestrejších vstupoch. Využite pochopenie problému (napríklad v jednoduchých špeciálnych prípadoch), teoretické vedomosti a nadhľad na overenie správnosti výstupu. Zadania riešte samostatne; budem penalizovať každé také riešenie, ktoré sa nápadne podobá na riešenie iného študenta.

O Vašich riešeniach sa spoločne porozprávame počas skúšky. Pri hodnotení budem zohľadňovať nasledovné kritériá: vhodnosť použitej metódy, správnosť, efektívnosť a originalita riešenia, zaradenie úlohy do širšieho kontextu, schopnosť modifikovať program podľa požiadaviek, schopnosť reagovať na teoretické otázky týkajúce sa danej úlohy, tvorivé nápady týkajúce sa možností vylepšenia a rozšírenia riešenia.

Za riešenie zadanej úlohy môžete získať maximálne 30 bodov, za odpoveď na teoretickú otázku maximálne 20 bodov. Celkové hodnotenie predmetu: $E = [25, 30)$, $D = [30, 35)$, $C = [35, 40)$, $B = [40, 45)$, $A = [45, 50]$.

Ak by ste mali akéhoľvek otázky, neváhajte mi napísať na e-mail `harman@fmph.uniba.sk`.

Šírenie požiaru

Najprv si prečítajte popis modelu horenia lesa na stránke Wikipedie. Uvažujme dvojrozmerný model ($d = 2$) na mriežke so stranou veľkosti L buniek. Pre bunku so súradnicami (x, y) berieme ako susedné bunky $(x - 1, y)$, ak $x > 1$, $(x + 1, y)$, ak $x < L$, $(x, y - 1)$, ak $y > 1$ a $(x, y + 1)$, ak $y < L$. Každá bunka môže byť v stave "prázdna", "strom" a "horiaci strom". Pravidlá zmien stavu buniek sa vykonajú v každom kroku simultánne, čiže:

1. Ak je v kroku i nejaká bunka v stave "horiaci strom", tak v kroku $i + 1$ bude táto bunka v stave "prázdna".

2. Ak je v kroku i nejaká bunka v stave "strom" a aspoň jedna jej susedná bunka je v kroku i v stave "horiaci strom", tak v kroku $i+1$ bude táto bunka v stave "horiaci strom".

3. Ak je v kroku i nejaká bunka v stave "strom" a ani jedna jej susedná bunka nie je v kroku i v stave "horiaci strom", tak v kroku $i+1$ bude táto bunka v stave "horiaci strom" s pravdepodobnosťou f (a s pravdepodobnosťou $1-f$ v stave "strom")[1].

4. Ak je v kroku i nejaká bunka v stave "prázdna", tak v kroku $i+1$ bude táto bunka v stave "strom" s pravdepodobnosťou p (a s pravdepodobnosťou $1-p$ v stave "prázdna").

Predpokladáme, že v prvom kroku sú všetky bunky v stave "prázdna".

Napíšte program so vstupnými hodnotami f, p, L, N, Δ kde f, p, L sú parametre s významom popísaným vyššie a N je počet krokov v jednom simulačnom behu (Δ je definované nižšie). Výsledkom programu bude interval spoľahlivosti pre počet r buniek v stave "strom" po N-tom kroku algoritmu. Program opakuje simulačné behy až dovtedy, pokiaľ nie je dĺžka 95-percentného intervalu spoľahlivosti pre r kratšia ako Δ. Po každom simulačnom behu program graficky znázorní stav buniek v mriežke.

Simulačné výsledky interpretujte. Prémiové body udelím za akékoľvek zmysluplné vylepšenie modelu (a programu), pokiaľ bude reálnejšie odrážať priebeh požiaru lesa.

Telekomunikačná sieť

Najprv si prečítajte sprievodný text na internetovej stránke: `http://plus.maths.org/content/call-routing-telephone-networks`.

Uvažujme teda nasledovný jednoduchý komunikačný model: Máme 5 uzlov (miest) A,B,C,D,E, pričom medzi každými dvomi mestami existuje priame telefonické spojenie s kapacitou $c \in \mathbb{N}$ simultánnych hovorov. Požiadavky na hovory vznikajú navzájom nezávisle, s intenzitou konštantnou v čase, rovnomerne na množine všetkých dvojíc miest

$$\{A, B\}, \{A, C\}, \{A, D\}, \{A, E\}, \{B, C\}, \{B, D\}, \{B, E\}, \{C, D\}, \{C, E\}, \{D, E\}.$$

To znamená, že postupnosť $0 = t_0$, t_1, t_2,... okamihov vzniku požiadaviek na hovory je taká, že časové diferencie $t_i - t_{i-1}$ sú nezávislé náhodné premenné s exponenciálnym rozdelením so strednou hodnotou $\lambda > 0$ minút a požadované spojenia sa generujú rovnomerne náhodne na množine všetkých desiatich dvojíc miest. Predpokladáme, že trvanie každého hovoru má exponenciálne rozdelenie so strednou hodnotou $\mu > 0$ minút. Doby trvania sú navzájom nezávislé (a taktiež sú nezávislé na časoch t_1, t_2, ...).

V čase 0 má každé spojenie $\{X, Y\}$ priradený nejaký pomocný uzol Z (pričom $Z \neq X$, $Z \neq Y$). V okamihu keď vznikne požiadavka na spojenie $\{X, Y\}$, systém sa správa nasledovne. Ak kapacita spojenia $\{X, Y\}$ ešte nie je nasýtená, tak sa hovor spojí priamo (a tento hovor potrvá až do konca; nebude za žiadnych okolností predčasne prerušený). Ak je kapacita spojenia $\{X, Y\}$ nasýtená, čiže práve prebieha c hovorov medzi X a Y, systém vyskúša prepojiť požiadavku cez svoj momentálne pridelený pomocný uzol Z. Ak

[1] To môže modelovať vznietenie stromu úderom blesku.

sú súčasne voľné kapacity spojení $\{X, Z\}$ aj $\{Y, Z\}$, tak sa hovor prepojí cez Z a uzol Z zostáva pridelený ako pomocný prepojovací uzol pre spojenie $\{X, Y\}$ (pochopiteľne o jednotku za zníži zostávajúca kapacita spojení $\{X, Z\}$ aj $\{Y, Z\}$). Ak nie je voľné spojenie $\{X, Z\}$, alebo $\{Y, Z\}$, hovor sa odmietne (systém sa nepokúša použiť na prepojenie hovoru ďalší uzol) a spojeniu $\{X, Y\}$ sa náhodne priradí nový pomocný uzol, iný než X,Y a Z. Správanie sa systému budeme sledovať po dobu Δ minút.

Napíšte program, ktorého vstupom budú konštanty c, λ, μ, Δ, N a výstupom bude 95-percentný interval spoľahlivosti pre podiel odmietnutých požiadaviek na prepojenie ku všetkým vzniknutým požiadavkám na prepojenie (N je počet simulačných behov, čiže koľkokrát sa má simulovať daný časový úsek činnosti systému dĺžky Δ). Bonusové body získa ten, kto túto úlohu zmysluplným a originálnym spôsobom modifikuje (zovšeobecní, rozšíri) a danú modifikáciu správne vyrieši.

Hromadná obsluha

Máme systém hromadnej obsluhy s jedinou linkou[2], pričom doba obsluhy jedného zákazníka trvá konštantne λ minút a celkový denný časový interval, počas ktorého systém akceptuje zákazníkov, je $[0, \Delta]$ minút[3]. Zákazníci sa správajú navzájom nezávisle, pričom:

- Doby medzi príchodmi jednotlivých zákazníkov sú nezávislé a majú gamma rozdelenie so strednou hodnotou δ a koeficientom variácie ν. (Koeficient variácie je pomer smerodajnej odchýlky a strednej hodnoty.)

- Zákazník, ktorý vstúpi do systému, sa rozhodne zaradiť sa do fronty s pravdepodobnosťou $\exp(-Hn)$, kde n je počet zákazníkov vo fronte v okamihu vstupu tohto zákazníka.

- Ak sa zákazník nerozhodne zaradiť sa do fronty, okamžite opúšta systém a ak sa zákazník *rozhodne* zaradiť sa do fronty, tak už istotne počká na obsluhu.

Simulačne skonštruujte intervaly spoľahlivosti pre stredné hodnoty nasledovných náhodných charakteristík: a) Počet zákazníkov obslúžených počas jedného dňa, b) Maximálna dĺžka fronty počas jedného dňa, c) Jednodňový priemer čakania tých zákazníkov, ktorí sa rozhodli zaradiť sa do fronty.

Počet simulačných behov (simulovaných dní) voľte $n = 1000$. Parametre určujúce dynamiku systému a zákazníkov voľte: $\lambda = 0.5, 1, 2$ minúty, $\Delta = 600$ minút, $\delta = 1$ minúta, $\nu = 0.5, 1, 2$ a $H = \ln(2)/10$. (Spolu je to 9 kombinácií parametrov, t.j. výsledných intervalov spoľahlivosti.) Výsledky pre jednotlivé kombinácie parametrov porovnajte a slovne interpretujte. Na generovanie náhodných realizácií z gama rozdelenia použite funkciu `rgamma`.

Problematika spadá do oblasti simulovania systémov s diskrétnymi udalosťami: `https://en.wikipedia.org/wiki/Discrete-event_simulation`.

[2] Obsluha je typu FIFO - first in, first out.

[3] Po čase Δ už systém neakceptuje nových zákazníkov, ale doobsluhuje zákazníkov vo fronte.

Fragmentácia názorov

Najprv si prečítajte článok https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0213246 Na základe popisu modelu v časti "The model" vytvorte program, ktorý bude generovať grafy z Figure 5 (pre zadané hodnoty γ a ϵ ako vstupné parametere). Výsledky pre jednotlivé kombinácie parametrov porovnajte a slovne interpretujte. Napíšte tiež program, ktorý pre zadané γ a ϵ a počet simulačných behov $n = 100$ vypočíta približný 95-percentný interval spoľahlivosti pre čas, ktorý uplynie po plnú homogenizáciu názorových zhlukov (čiže po okamih, po ktorom už žiadny člen populácie nemôže v rámci predpokladov modelu zmeniť názor).

Nekorektné testovanie

V poslednej dobe sa veľa diskutuje o kríze reprodukovateľnosti empirických vedeckých zistení. Jedným z dôvodov je enormný tlak na čo najvyšší počet "zaujímavých" publikácií, čo sa niektorí menej čestní vedci snažia dosiahnuť porušením základných princípov testovania štatistických hypotéz.

Predpokladajme pre jednoduchosť klasický jednovýberový t-test hypotézy $H_0 : \mu = \mu_0$ voči $H_1 : \mu \neq \mu_0$, kde μ je skutočná stredná hodnota normálneho náhodného výberu a μ_0 je testovaná stredná hodnota. Ako vieme zo základného kurzu, štatistika $T_n = \sqrt{n}\frac{\bar{X}_n - \mu_0}{S_n}$ pre tento test má za podmienky platnosti H_0 Studentovo rozdelenie t_{n-1}, kde n je rozsah náhodného výberu, $\bar{X}_n$ je jeho aritmetický priemer a S_n je jeho výberová smerodajná odchýlka.

Korektný prístup k použitiu t-testu na hladine významnosti 0,05 je stanoviť si vopred rozsah náhodného výberu n, vykonať presne n meraní/pozorovaní a zo všetkých dát vypočítať testovaciu štatistiku T_n. Ak bude T_n v kritickej oblasti testu, čiže ak bude hodnota $|T_n|$ väčšia ako 97,5-percentný kvantil rozdelenia t_{n-1}, tak H_0 zamietneme, inak H_0 nezamietneme. Takto vykonávaný test má chybu prvého druhu presne 0,05. Čiže, ak naozaj platí $\mu = \mu_0$, tak tento fakt zamietneme len s pravdepodobnosťou 0,05.

Obvykle je však experiment nastavený tak, že nezamietnutie hypotézy H_0 znižuje šancu, že daný výskum bude publikovaný (H_0 reprezentuje "nudný" výsledok a H_1 reprezentuje "atraktívny" výsledok.) Relatívne častý nekorektný spôsob štatistického testovania je preto tento:

Výskumník si stanoví minimálny rozsah $n_{\min}$ a maximálny rozsah $n_{\max}$ náhodného výberu. Výskumník najprv vykoná $n_{\min}$ pozorovaní a skontroluje, či t-štatistika vypočítaná z dostupných pozorovaní nie je v kritickej oblasti t-testu (pre $n = n_{\min}$). Ak je v kritickej oblasti, výskumník ukončí experiment s výsledkom "H_0 zamietame". Ak však nie je v kritickej oblasti, výskumník vykoná ďalšie pozorovanie a opäť skontroluje, či t-štatistika vypočítaná z dostupných pozorovaní nie je v kritickej oblasti t-testu (teraz už pre $n = n_{\min} + 1$). Ak je v kritickej oblasti, výskumník ukončí experiment s výsledkom "H_0 zamietame". Ak nie je v kritickej oblasti, výskumník vykoná ďalšie pozorovanie a tak ďalej. Jedine v prípade, že ani po $n_{\max}$-tom pozorovaní nepadne t-štatistika do kritickej oblasti, výskumník ukončí experiment s výsledkom "H_0 nezamietame".

Vašou úlohou v tomto príklade je napísať program, ktorý simuluje použitie t-testu vyššie uvedeným spôsobom a vypíše 95-percentný interval spoľahlivosti pre *skutočnú* chybu prvého druhu pri tomto spôsobe testovania hypotézy H_0. Ako vstup programu voľte hodnoty μ, σ^2, $n_{\min}$, $n_{\max}$, N, kde μ je skutočná stredná hodnota pozorovaní, σ^2 je skutočná disperzia pozorovaní, $n_{\min} \leq n_{\max}$ sú ohraničenia na veľkosť výberu spomenuté vyššie a N je počet nezávislých simulácií celého experimentu.

Pozitívne hodnotím akékoľvek bystré zjednodušenia, rozšírenia a komentáre relevantné k tomuto problému. Nedávno tiež vyšiel na túto tému nasledovný vedecký článok, ktorý si môžete prečítať na hlbšie pochopenie celej problematiky: `https://www.biorxiv.org/content/10.1101/2019.12.12.868489v1`.

Vaše vlastné zadanie

Môžete si vymyslieť aj vlastné zadanie, je však potrebné, aby ste mi ho poslali na schválenie aspoň dva týždne pred skúškou. Môže to byť stochastická simulácia čohokoľvek, s čím ste sa stretli a čo Vás zaujalo, je podobnej zložitosti ako zadania vyššie, nemalo by to však byť súčasťou Vašej diplomovej práce či zamestnania. Námety: Simulovanie šírenia sa epidémie (množstvo existujúcich modelov), simulovanie stratégie parkovania (napríklad `https://arxiv.org/abs/1904.06612`), stochastická simulácia biochemických reakcií (napríklad `https://royalsocietypublishing.org/doi/10.1098/rsif.2018.0943`) a tak ďalej.